Tracing Bilingual Knowledge Memorization and Generalization Dynamics in Domain Adaptation

Xin Zhao[1,3] Naoki Yoshinaga[2,3] Yuma Tsuta[3]* Akiko Aizawa[3]
[1]The University of Tokyo [2]Institute of Industrial Science, The University of Tokyo
[3]Research and Development Center for Large Language Models, National Institute of Informatics
xzhao@tkl.iis.u-tokyo.ac.jp ynaga@iis.u-tokyo.ac.jp aizawa@nii.ac.jp

Abstract

Our study explores how domain knowledge is learned within a language and transferred across languages via multilingual domain adaptation (ML-DA) of language models. We first propose **AdaXEval**, an adaptive evaluation method that builds multiple-choice QA datasets from the same bilingual domain corpus used for training, enabling a direct and controlled analysis of multilingual knowledge acquisition. Through continual training experiments, we further uncover a **Loss-shielding** phenomenon, showing how LLMs memorize and generalize facts during training, and we reveal the mechanism of how domain training transform data into internalized knowledge. MLDA-Eval

1 Introduction

Since large language models (LLMs) are likely to cover knowledge that commonly appears in the training data, it often struggles in specialized domains with rare knowledge [1]. Domain adaptation addresses this by continually training LLMs on domain-specific data to enhance expertise [2, 3]. However, achieving efficient knowledge acquisition in low-resource settings remains challenging, motivating a deeper understanding of the domain adaptation process. Prior studies show that facts accumulates gradually through repeated exposures, shaped by data frequency, model scale [4, 5]. However, Most analyses focus on general relational facts rather than domain facts and the link between data and knowledge remains underexplored.

Our work aims to investigate the process of knowledge acquisition in domain adaptation from a interpretability perspective. Specifically, we seek to understand, during the continual-training on domain corpora, how domain facts are *memorized* and *generalized* across different linguistic contexts, including both *intralingual* (within a language) and *interlingual* (across languages) variations, and to identify factors facilitating effective knowledge memorization and transfer. To achieve the goal, we ask: **(RQ1)** how domain knowledge acquisition can be effectively evaluated? and **(RQ2)** what is the mechanism behind the transformation from training data to knowledge?

Existing domain adaptation evaluation primarily rely on public benchmarks [6, 7] or training loss analysis [8, 9, 5]. However, such benchmarks offer limited coverage for low-resource domains or languages, and fail to capture knowledge generalization abilities. Moreover, the misalignment between training data and benchmark knowledge coverage makes evaluation an imperfect reflection of acquired knowledge. To address these gaps and resolve **RQ1**, we propose **AdaXEval**, an adaptive domain knowledge evaluation pipeline. It automatically constructs multiple-choice datasets to evaluate knowledge **memorization**, **intralingual, and intralingual generalization**.

We next investigate how training data is dynamically transformed into knowledge (**RQ2**). We conduct a case study of Japanese biomedicine domain adaptation using a 13B English/Japanese bilingual LLM [10]. We begin with monolingual continual training on both English and Japanese using the J-STAGE corpus, which contains biomedical documents for both languages. By evaluating training checkpoints with AdaXEval-generated datasets, we observe a gradual knowledge memorization and intralingual generalization; however, LLM struggles to achieve cross-lingual transfer. Our analysis reveals that knowledge is memorized and generalized as losses of correct options are shielded from loss growth due to model's overfitting to training data, which we term **loss shielding**.

[*] Currently, he works for Fixstars, Inc.

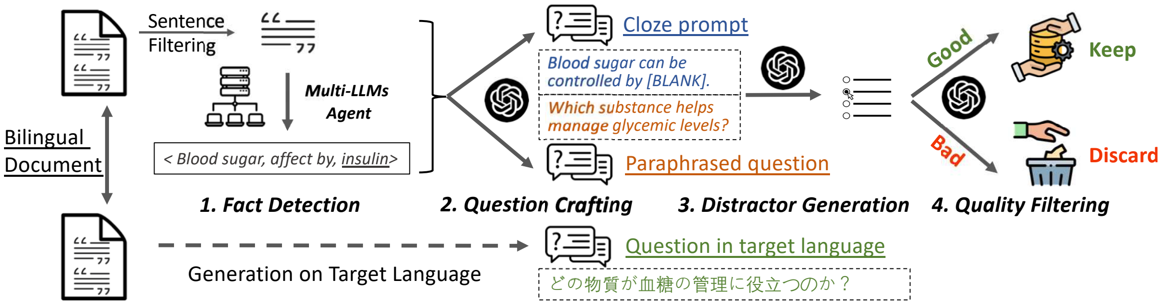

Bilingual Document
Sentence Filtering
Multi-LLMs Agent
< Blood sugar, affect by, insulin>
1. Fact Detection
Cloze prompt
Blood sugar can be controlled by [BLANK].
Which substance helps manage glycemic levels?
Paraphrased question
2. Question Crafting
3. Distractor Generation
Good
Bad
Keep
Discard
4. Quality Filtering
Generation on Target Language
Question in target language
どの物質が血糖の管理に役立つのか？

Figure 1: Overview of AdaXEval, a pipeline to adaptively generate domain knowledge evaluation datasets.

2 Domain Adaptation Evaluation

2.1 AdaXEval Pipeline

AdaXEval is an adaptive domain knowledge evaluation pipeline. It evaluates knowledge acquisition by generating evaluation datasets directly from training corpora, ensuring evaluated facts stay aligned with training data (Figure 1).

Fact Detection AdaXEval first detects sentences that may contain domain facts through a two-step strategy: named-entity-recognization (NER)-based sentence filtering and multi-agent fact detection. First, domain-specific NER tools and linguistic heuristics are employed to identify sentences in the training corpora that contain multiple named entities. Next, we design Chain-of-Thought (CoT) instructions to detect sentences containing domain facts from the filtered sentences, and extract triples in the format ⟨`subject, relation, object`⟩ as the reference for question crafting. Specifically, AdaXEval employs a multi-LLM agent for fact detection and triple extraction.

Queries Crafting Given the factual sentence and referenced triple, we first prompt the LLM to generate reliable domain factual triples (*e.g.*, ⟨*blood sugar level, can be controlled by, insulin*⟩). While subjects and objects are preferably named entities, descriptive phrases are acceptable. AdaXEval uses advanced LLMs to generate diverse question-answer pairs measuring three abilities:

1) Knowledge memorization uses a cloze prompt with `[BLANK]` as the placeholder for the object (*e.g.*, *Blood sugar level can be controlled by [BLANK].*). Given the original sentence, we prompt the LLM to generate a cloze question that closely matches the original sentence.

2) Intralingual generalization assesses LLMs' ability to acquire knowledge using linguistic expressions that vary from those in the training corpus. We design CoT instructions to let LLMs paraphrase the cloze queries into question-like style questions (*e.g.*, `Which substance helps manage glycemic levels in the body?`)

3) Interlingual generalization measures how learned facts can be transferred across languages. We adapt AdaXEval to a bilingual domain corpus containing languages X and Y, using the paraphrased dataset from language X to evaluate cross-lingual transfer capabilities in Y.

Distractor Generation AdaXEval then generates three plausible yet incorrect answer options that remain topically related but unambiguously wrong, while explicitly instructing the advanced LLM to avoid surface-level cues such as sequence length.

Quality Filtering Finally, AdaXEval uses the LLM to filter low-quality multiple-choice QA instances that fail to meet the requirements in § 2.1.

Evaluation metric For each evaluation dataset, we follow [11] to compute the average cross-entropy loss over the target tokens of possible answers and select the one with the highest generation possibility as the final answer. We report prediction accuracy as the metric. See Appendix A for the mathematical formulation of the evaluation metric.

2.2 Construction of AdaXEval Datasets

Domain corpus: Our study investigates biomedical domain adaptation in English–Japanese as a case study. Specifically, we utilize the J-STAGE, an English-Japanese bilingual biomedical corpus (see § 3.1), as the data source for both model training and AdaXEval generation.

Details of Generation: We randomly sampled 10,000 bilingual documents to generate the evaluation dataset. We split abstracts into sentences and filter out sentences with fewer than two biomedical entities. For fact detection of filtered sentences, we ensemble generation from three LLMs from different families to improve fact detection reliability.

Table 1: Dataset statistics at each step of AdaXEval.

Step	English	Japanese
Sampled abstracts	10,000	10,000
Splitted sentences	81,770	71,661
Sentences after entity filtering	45,390	40,762
Triple extraction	4840	3926
Cloze queries / Paraphrases	4840	3926
After Quality Filtering	3236	2553

Finally, we use GPT-4.1 to generate cloze queries, paraphrases, and three distractors for each instance. The statistical report of generated datasets is shown in Table 2. We also conduct human evaluation and the evaluation result indicates that AdaXEval is able to generate high-quality evaluation data (See more in Appendix B).

3 Tracing Knowledge Acquisition

This section examines the training dynamics of domain adaptation and explores the mechanism underlying the transformation from training data to knowledge in the monolingual setting. We do monolingual continual training on English and Japanese using biomedical datasets.

3.1 Experimental Setup

Data preparation: We utilize a subset of the J-STAGE corpus, which comprises Japanese research papers with some abstracts translated into English.[1] Specifically, we select 614,444 Japanese and 404,643 English biomedical documents, paired one-to-one. These bilingual pairs cover source data for AdaXEval generation. To strengthen domain adaptation and enable fine-grained analysis, we apply instruction pretraining as a data augmentation method [12]. The raw documents are then combined with the generated instructions for continual training.

Training setup: We adopt llm-jp-3-13B [10], a strong Japanese–English bilingual LLM, as the base model for pretraining, owing to its superior language ability in both languages, particularly Japanese. For each language, we cut off 0.5B tokens from the constructed corpus and train the data on llm-jp-3-13B for four epochs.

3.2 Tracing Performance Dynamics

Figure 2 reports AdaXEval evaluation results, indicating that domain knowledge is gradually acquired during training in both English and Japanese monolingual training.

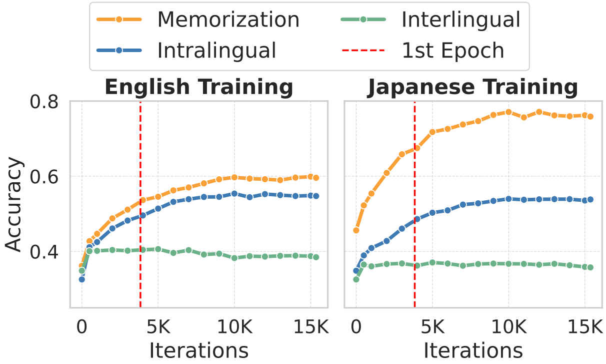


Figure 2: Dynamic knowledge acquisition evaluation

Memorization evaluation: Accuracy increases from 36.1% to 59.6% (**+23.5%**) in English and from 45.7% to 75.9% (**+30.2%**) in Japanese. The higher post-training accuracy in Japanese partly reflects the stronger medical knowledge base of llm-jp-3-13B. However, since the factual instances used for evaluation differ between languages, direct cross-lingual comparison is not strictly fair.

Intralingual generalization: Both languages exhibit strong performance on the paraphrased datasets, with accuracy increasing from 32.6% to 54.7% (**+22.1%**) in English and from 34.9% to 53.8% (**+18.9%**) in Japanese. Notably, the improvement in English paraphrases parallels the memorization gain, whereas it is about 10% lower in Japanese, suggesting that the difficulty of intralingual generalization differs across languages.

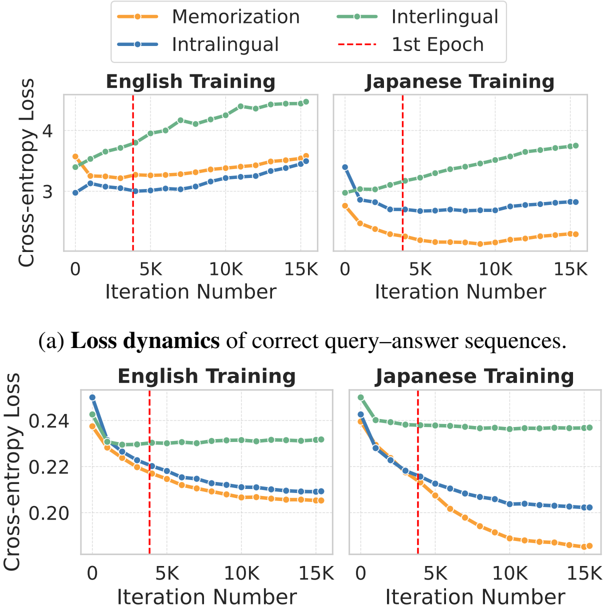
Interlingual generalization: Figure 2 reveals that monolingual training results in limited cross-lingual knowledge transfer, yielding only **3.6%** improvement in English-to-Japanese and **3.1%** improvement in reverse direction.

4 Acquiring Fact via Loss Shielding

This section analyzes the sequence loss of evaluation data to understand how knowledge is memorized and generalized during training. We analyze loss as it directly drives predictions on our multiple-choice datasets (see § 2.1) and reflects the model's generation behavior, where sequences with lower loss are more likely to be generated.

(1) Training overfits to data, but the loss shielding drives knowledge memorization. We first calculate the sequence loss of queries paired with correct answers. Figure 3a shows the loss trajectory across training checkpoints for English and Japanese training. On the cloze prompt dataset, the loss decreases in early training but rises in

[1] Access to the dataset is restricted by the J-STAGE license, so it cannot be publicly released.



(a) **Loss dynamics** of correct query–answer sequences.

(b) **Loss ratio dynamics** for sequences with correct answers relative to all candidates.

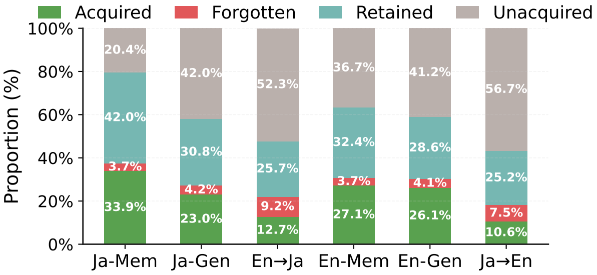


Figure 4: Instance state transitions before/after training.

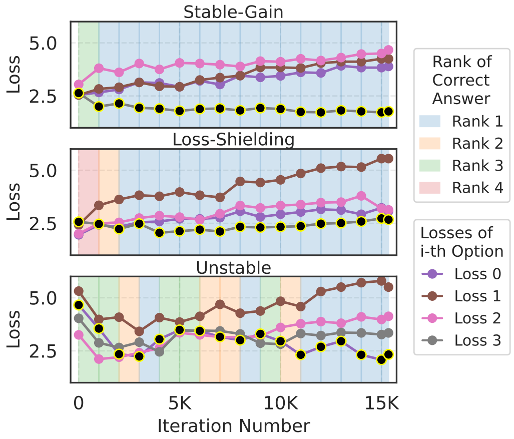


Figure 5: Acquired instances with different loss shapes. The line with bright circles indicates the correct answer.

later training, suggesting that training causes the model to overfit to the training corpus. However, Figure 2 reveals that memorization accuracy continues to improve until the 10K-th iteration. To investigate this divergence, we then measure the ratio of the correct-sequence loss to the total loss across all four options. As shown in Figure 3b, this ratio mirrors the accuracy trend, suggesting that knowledge can still be memorized even under overfitting, since correct sequences are shielded from rapid loss growth, a phenomenon we term *loss shielding*. Figure 3b also explains the gap between cloze prompts and paraphrases, which the significant loss ratio gap in Japanese predicts, as shown in Figure 2.

(2) A trade-off exists between knowledge acquisition and forgetting. For each instance, we check its state transition before and after training. Figure 4 shows the proportions of instances which knowledge is retained, acquired, forgotten, or unacquired during training. Forgetting remains limited in monolingual evaluations, including both memorization and intralingual generalization. In contrast, cross-lingual transfer exhibits a notable increase in forgotten cases, offsetting gains from newly acquired knowledge. This suggests that while training monolingually can introduce transferable knowledge, it also causes a decline in performance in other languages due to forgetting.

(3) Instance-level case studies show diverse knowledge acquisition patterns. We then examine the loss dynamics of instances acquired after training, analyzing all four options. We observe that the loss dynamics of correct answers follow distinct patterns: they either decrease steadily (Stable-Gain), increase while remaining lower than incorrect options (Loss-Shielding), or exhibit unstable behavior. Figure 5 illustrates three examples corresponding to the three loss change patterns.

5 Conclusion

In this paper, we investigated how LLMs acquire domain knowledge and transfer it across languages. We proposed **AdaXEval**, an adaptive evaluation pipeline that automatically generates datasets to assess memorization, intralingual generalization, and cross-lingual transfer of domain knowledge. Using benchmark generated from J-STAGE bilingual corpus by AdaXEval, we conduct a case study focusing on English-Japanese biomedical domain adaptation. We analyze the training dynamics of domain adaptation and find that knowledge acquisition is driven by **loss shielding**, wherein overfitting raises losses on irrelevant representations more rapidly than on relevant ones.

References

[1]Joel Jang, Seonghyeon Ye, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim, Stanley Jungkyu Choi, and Minjoon Seo. Towards continual knowledge learning of language models, 2022.

[2]Junfeng Jiang, Fei Cheng, and Akiko Aizawa. Improving referring ability for biomedical language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Findings of the Association for Computational Linguistics: EMNLP 2024**, pp. 6444–6457, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[3]Mathieu Laï-king and Patrick Paroubek. Pre-training data selection for biomedical domain adaptation using journal impact metrics, 2024.

[4]Xin Zhao, Naoki Yoshinaga, and Daisuke Oba. Tracing the roots of facts in multilingual language models: Independent, shared, and transferred knowledge. In Yvette Graham and Matthew Purver, editors, **Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 2088–2102, St. Julian's, Malta, March 2024. Association for Computational Linguistics.

[5]Yihong Liu, Mingyang Wang, Amir Hossein Kargaran, Felicia Körner, Ercong Nie, Barbara Plank, François Yvon, and Hinrich Schütze. Tracing multilingual factual knowledge acquisition in pretraining, 2025.

[6]Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Aguera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge, 2022.

[7]Junfeng Jiang, Jiahao Huang, and Akiko Aizawa. JMedBench: A benchmark for evaluating Japanese biomedical large language models. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, **Proceedings of the 31st International Conference on Computational Linguistics**, pp. 5918–5935, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics.

[8]Hoyeon Chang, Jinho Park, Seonghyeon Ye, Sohee Yang, Youngkyung Seo, Du-Seong Chang, and Minjoon Seo. How do large language models acquire factual knowledge during pretraining?, 2024.

[9]Nicolas Zucchet, Jörg Bornschein, Stephanie Chan, Andrew Lampinen, Razvan Pascanu, and Soham De. How do language models learn facts? dynamics, curricula and hallucinations, 2025.

[10]LLM-jp, :, Akiko Aizawa, Eiji Aramaki, Bowen Chen, Fei Cheng, Hiroyuki Deguchi, Rintaro Enomoto, Kazuki Fujii, Kensuke Fukumoto, Takuya Fukushima, Namgi Han, et al. Llm-jp: A cross-organizational project for the research and development of fully open japanese llms, 2024.

[11]Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, 12 2023.

[12]Ting Jiang, Shaohan Huang, Shengyue Luo, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. Improving domain adaptation through extended-text reading comprehension, 2024.

A Evaluation Metrics

We follow [11] to compute the average cross-entropy loss over the target tokens of possible answers and select the one that has the highest generation possibility as the final answer. Specifically, for loss calculation of cloze queries, we use tokens before the `[BLANK]` as context and compute loss on the following tokens. For paraphrases, we treat the question as context and measure only the loss of answer tokens. We use prediction accuracy as the evaluation metric for knowledge acquisition.

Formulation: Let the model be $p_\theta(\cdot \mid \cdot)$. Each dataset $\mathcal{D}$ contains pairs (q, a), where q is either a **cloze prompt** or a **paraphrase question**, and $a = (a_1, \ldots, a_m)$ is the tokenized answer sequence (*e.g.*, "insulin"). For cloze prompts, the prompt contains a special token `[BLANK]`, and for paraphrases, the question is a natural question. We denote by c the context tokens and by $s = (s_1, \ldots, s_n)$ the evaluation sequence whose loss we measure.

Cloze queries: For a cloze prompt such as:
`"[BLANK] can be used to control blood sugar level."`
the evaluation sequence s is the full completion after the `[BLANK]`, i.e.,

$$s = (a_1, \ldots, a_m, r_1, \ldots, r_k),$$

where $a_1, \ldots, a_m$ are answer tokens ("insulin"), and $r_1, \ldots, r_k$ are the remainder tokens (" can be used to control blood sugar level"). The average cross-entropy loss is defined as

$$\mathcal{L}_{\text{cloze}}(q, a) = -\frac{1}{n} \sum_{t=1}^{n} \log p_\theta(s_t \mid c, s_{<t}).$$

Paraphrase queries: For a paraphrase question such as `"Which substance helps manage glycemic levels?"` we take the entire question tokens as context c, and the evaluation sequence is only the answer tokens:

$$s = (a_1, \ldots, a_m).$$

The loss is computed as

$$\mathcal{L}_{\text{para}}(q, a) = -\frac{1}{m} \sum_{t=1}^{m} \log p_\theta(a_t \mid c, a_{<t}).$$

Prediction and accuracy: For multiple-choice answers $\mathcal{A} = \{a^{(1)}, \ldots, a^{(K)}\}$, we select the candidate with the lowest loss (equivalently, highest likelihood):

$$\hat{a} = \arg\min_{a \in \mathcal{A}} \mathcal{L}(q, a).$$

Finally, the knowledge acquisition metric is accuracy:

$$\text{Accuracy} = \frac{1}{|\mathcal{D}|} \sum_{(q,a) \in \mathcal{D}} \mathbf{1}[\hat{a} = a].$$

Table 2: Human annotation of AdaXEVal datasets generated from the biomedical J-STAGE corpus.

Evaluation Metric	Japanese	English
Cloze prompt (Faithfulness)	2.84/3	2.89/3
Paraphrase (Fluency)	2.94/3	2.96/3
Paraphrase (Diversity)	2.48/3	2.62/3
Paraphrase (Factuality)	2.86/3	2.86/3
Paraphrase (Inter-Factuality)	1.68/2	1.76/2
Distractor (Plausibility)	2.35/3	2.16/3
Distractor (Incorrectness)	2.89/3	2.97/3

B Human Annotation

To assess the quality of our generated datasets, we conduct a comprehensive human evaluation across four key components of the knowledge extraction and question generation pipeline. The annotation is conducted by the first author, who has a language background of both Japanese and English. For the annotation that requires specific domain knowledge, the author uses advanced LLMs, such as ChatGPT or Claude, as an assistant for annotation.

(1) Cloze prompt evaluation checks the faithfulness of the generated prompt to the original sentence structure.

(2) Paraphrases evaluation is conducted on four dimensions: fluency and grammaticality, linguistic diversity in reformulation, factual correctness to the original sentence in the source language, and interlingual factual correctness using the corresponding documents in the target language.

(3) Distractor quality is measured through plausibility within the domain and apparent incorrectness relative to the original context.

Each metric employs structured scoring rubrics with scales ranging from 0-2 or 0-3 following a carefully designed humanin evaluation guideline. We randomly sample 50 instances for each language and conduct human evaluation following the guideline. As shown in Table 2, our evaluation results indicate that AdaXEval is capable of generating high-quality evaluation data, meeting the requirements for assessing knowledge memorization as well as intralingual and interlingual generalization evaluation.